Bayesian Analysis (2019) **TBA**, Number TBA, pp. 1–32

Efficient Bayesian Model Choice for Partially Observed Processes: With Application to an Experimental Transmission Study of an Infectious Disease

Trevelyan J. McKinley*, Peter Neal†, Simon E. F. Spencer‡, Andrew J. K. Conlan§, and Laurence Tiley§

Abstract. Infectious diseases such as avian influenza pose a global threat to human health. Mathematical and statistical models can provide key insights into the mechanisms that underlie the spread and persistence of infectious diseases, though their utility is linked to the ability to adequately calibrate these models to observed data. Performing robust inference for these systems is challenging. The fact that the underlying models exhibit complex non-linear dynamics, coupled with practical constraints to observing key epidemiological events such as transmission, requires the use of inference techniques that are able to numerically integrate over multiple hidden states and/or infer missing information. Simulation-based inference techniques such as Approximate Bayesian Computation (ABC) have shown great promise in this area, since they rely on the development of suitable simulation models, which are often easier to code and generalise than routines that require evaluations of an intractable likelihood function. In this manuscript we make some contributions towards improving the efficiency of ABC-based particle Markov chain Monte Carlo methods, and show the utility of these approaches for performing both model inference and model comparison in a Bayesian framework. We illustrate these approaches on both simulated data, as well as real data from an experimental transmission study of highly pathogenic avian influenza in genetically modified chickens.

Keywords: Bayesian model choice, infectious disease models, partially observed processes, particle MCMC, Approximate Bayesian Computation.

1 Introduction

Infectious diseases such as avian influenza pose a global threat to human health. Mathematical modelling can elucidate on key mechanisms that underlie disease spread, which can directly inform the control of potential outbreaks. However, the utility of these approaches depends greatly on the ability to calibrate them to observed data. For ex-

*College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn, UK, t.mckinley@exeter.ac.uk

†Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

‡Department of Statistics and the Warwick Analytical Sciences Centre, University of Warwick, Coventry, UK

§Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

 https://doi.org/10.1214/19-BA1174

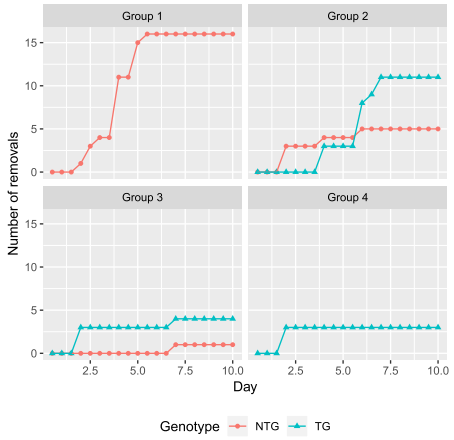


Figure 1: Data taken from Lyall et al. (2011).

Experiment	Exposure	Genotype	Number of birds
1	In-contact	NTG	12
	Challenge		5
2	In-contact	TG	12
	Challenge	NTG	5
3	In-contact	NTG	12
	Challenge	TG	5
4	In-contact	TG	12
	Challenge		5

Table 1: Experimental conditions (from Lyall et al., 2011).

ample, Lyall et al. (2011) genetically modified chickens to be resistant to avian influenza. These *transgenic* chickens expressed a short-hairpin RNA designed to function as a decoy, and hence interfere and hinder viral propagation. In order to assess the efficiacy of the modification they ran a series of natural transmission experiments. They used a *crossed* experimental design in which four experiments were undertaken, where each experiment used 5 *challenge* birds (that had been inoculated with highly pathogenic avian influenza—HPAI) co-housed with 12 uninfected *in-contact* birds. Each experiment was run over 10 days, with each bird being swabbed twice daily (producing buccal and cloacal swabs). Some birds were artificially removed from the study (if they were clinically sick, or sometimes for other blood work to be done), some birds died naturally, and others were euthanased at the end of the study. From these data it is possible to derive integer-valued time series counts for the number of infections for each genotype over each half-day period, which are shown in Figure 1.

It is clear from Figure 1 that there are strong differences between the genotypes, such that experiments in which the non-transgenic (NTG) birds constituted the challenge group resulted in many birds dying in greater numbers (and more quickly) than exper-

iments where the transgenic (TG) birds constituted the challenge group. In Lyall et al. (2011) these differences were compared using a series of Bonferroni corrected Mann-Whitney tests. However, a key question that remained was whether these patterns could be caused by differences in transmission potential between the genotypes, differences in susceptibility to infection, or both. The supposition in Lyall et al. (2011) was that it was the former, but the crude Mann-Whitney test is not sufficient to provide evidence for or against these competing hypotheses. In this paper we develop a framework to estimate the Bayesian evidence (e.g. Kass and Raftery, 1995) for a series of competing dynamic transmission models that allows us to directly address these questions.

2 Statistical methods

Infectious disease systems often have complex non-linear dynamics, which are not well captured by classical statistical approaches to inference, such as generalised linear modelling or null hypothesis significance testing. A common way to model these processes is to use compartmental models, where individuals progress through a series of different epidemiological states over time (e.g. Anderson and May, 1991). Stochastic versions of these models are parameterised by the average rates of progression through the compartments, which are dependent on the state of the system at any given time (e.g. Keeling and Rohani, 2008).

The main challenge for inference is that even in highly controlled settings it is not possible to directly observe key measurements required to reconstruct a likelihood function, such as the time of infection, or even the infection status of animals (particularly in presence of sub-clinical infections). Hence the problem becomes that of inference for hidden Markov (or non-Markovian) models (HMMs).

The Bayesian paradigm provides a natural framework within which to tackle these problems, since the uncertainties associated with the hidden process are propagated through the system and explicitly incorporated into the parameter estimates and predictions from the model. Here we wish to estimate the posterior distribution for the parameters, $\boldsymbol{\theta}$, given the observed data $\boldsymbol{y}$, which is in turn dependent on hidden states $\boldsymbol{x}$. In this case the posterior, $f(\boldsymbol{\theta} \mid \boldsymbol{y})$, is defined as:

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta})\, f(\boldsymbol{\theta}) = \left[\int_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})\, d\boldsymbol{x} \right] f(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\mathcal{X}}$ corresponds to the sample space for the hidden variables $\boldsymbol{x}$ and $f(\boldsymbol{\theta})$ is the *prior* distribution for the parameters $\boldsymbol{\theta}$. The idea is that by modelling the missing information, the joint distribution $f(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})$ has a specific (and known) mathematical form.

Nevertheless, in practice, the integral in (1) is often still analytically intractable, and thus to evaluate (1) it is necessary to *numerically* integrate over the sample space for the hidden states in a computationally feasible manner. Recent advances for addressing this problem involve replacing the calculation of $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ with a Monte Carlo (MC) estimate derived from repeated simulations from the underlying dynamic model (e.g. Tavaré et al., 1997; Marjoram et al., 2003; Beaumont, 2003; Sisson et al., 2007; Toni et al., 2009;

Andrieu and Roberts, 2009; Andrieu et al., 2010). Under certain conditions, Andrieu and Roberts (2009) showed that as long as the estimate of the likelihood is *unbiased* and *non-negative*, then a Markov chain Monte Carlo (MCMC; see e.g. Gilks et al., 1996) algorithm can be developed that produces samples from the correct posterior distribution in probability (see also Beaumont, 2003). Due to the dynamic nature of infectious disease models, direct simulation is often inefficient, but the structure of the models can often be exploited to produce more efficient estimators. For example, particle filtering methods (see e.g. Doucet et al., 2001) are now often used to produce a non-negative and unbiased likelihood estimate for state-space models, which can have a lower variance than vanilla MC estimators. These hybrid methods are known as *particle MCMC* (PMCMC; e.g. Andrieu et al., 2010; Drovandi et al., 2016; Alzahrani et al., 2018), and are attractive because it is often far easier to code a simulation model than it is to optimise more traditional approaches to dealing with HMMs, such as data-augmented (reversible-jump) MCMC (e.g. Gibson and Renshaw, 1998; O'Neill and Roberts, 1999; Jewell et al., 2009). Despite this, PMCMC methods can be highly computationally intensive, but in practice they can be made more efficient by introducing a discrepancy function, such that simulations do not have to be exactly consistent with the observed data but rather have to be within some region 'close' to the data (e.g. Del Moral et al., 2015; Drovandi et al., 2016). This reduces the computational burden of the routines at the cost of producing an approximate (rather than exact) posterior. Recently Wilkinson (2013) provides an alternative interpretation of this approximate distribution as the exact posterior distribution for a model incorporating specific model discrepancy (or measurement error) terms.

These latter approaches come under a generic suite of methods known as Approximate Bayesian Computation (ABC; e.g. Tavaré et al., 1997; Marjoram et al., 2003; Toni et al., 2009; McKinley et al., 2009, 2018), which aim to improve the efficiency of numerical estimation algorithms through both dimension reduction techniques (by fitting to summary measures of the data, rather than the full data), in addition to the use of discrepancy measures. However, in this manuscript we place a discrepancy function around *each data point*, and we do not further approximate by reducing the data to a set of *summary statistics*, which can cause challenges in ABC-based model choice routines (see e.g. Marin et al., 2014; though we note interesting recent work by Pudlo et al., 2016 and Raynal et al., 2017 that aim to perform robust model choice and inference respectively within the classic ABC paradigm using random forests—see e.g. Breiman, 2001). In our case, as the discrepancy tolerance reduces to zero around each data point, the approximate posteriors converge to the exact posterior (with no model discrepancy).

2.1 The alive particle filter

To fit the models we used the particle MCMC algorithm of Drovandi et al. (2016) using the alive particle filter (APF) of Del Moral et al. (2015). From now on we discuss the algorithms in the context of the systems presented in this paper, but note that other generalisations can also be used (e.g. Drovandi et al., 2016).

Consider data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ as a set of integer counts at time points $t = 1, \ldots, T$. These counts could be for different groups of individuals, and hence at time t the

data $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Gt})$ where $G \geq 1$. We also introduce a fixed set of *tolerances*, $\boldsymbol{\epsilon} = \{\boldsymbol{\epsilon}_t; \text{for } t = 1, \ldots, T\}$, where $\boldsymbol{\epsilon}_t = \{\epsilon_{gt}; \text{for } g = 1, \ldots, G\}$.

We can then condition the likelihood, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$, on the tolerances $\boldsymbol{\epsilon}$, and write this as:

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}) &= \int_{\boldsymbol{\mathcal{X}}} \left[f(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta}) \right] d\boldsymbol{x}, \\ &= \int_{\boldsymbol{\mathcal{X}}} \left[f_{X_0}(\boldsymbol{x}_0 \mid \boldsymbol{\theta}) \prod_{t=1}^{T} f_{Y_t|X_t}(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\epsilon}_t) f_{X_t|X_{t-1}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}) \right] d\boldsymbol{x}, \end{aligned} \tag{2}$$

where $f_{X_0}(\boldsymbol{x}_0 \mid \boldsymbol{\theta})$ is the density function for the initial conditions for the hidden states at time 0 (which may or may not depend on the unknown parameters $\boldsymbol{\theta}$). The density $f_{X_t|X_{t-1}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta})$ governs the progression from state $\boldsymbol{x}_{t-1}$ to state $\boldsymbol{x}_t$ based on the underlying model, and $f_{Y_t|X_t}(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\epsilon}_t)$ is the discrepancy/observation term for the data $\boldsymbol{y}_t$ conditional on the hidden states $\boldsymbol{x}_t$ and the tolerances $\boldsymbol{\epsilon}_t$. The integral in (2) is now over the multidimensional space $\boldsymbol{\mathcal{X}}$ corresponding to all possible values for the hidden states $\boldsymbol{x}_t$ at time points $t = 0, \ldots, T$. In this manuscript, we consider the discrepancy term, $f(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\epsilon}_t)$, to be an indicator function, such that

$$f(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\epsilon}_t) = \prod_{g=1}^{G} f_{Y_{gt}|X_{gt}}(y_{gt} \mid \boldsymbol{x}_{gt}, \epsilon_{gt}), \tag{3}$$

and

$$f_{Y_{gt}|X_{gt}}(y_{gt} \mid \boldsymbol{x}_{gt}, \epsilon_{gt}) = \begin{cases} 1 & \text{if } |h_g(\boldsymbol{x}_{gt}) - y_{gt}| \leq \epsilon_{gt}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here $h_g(\cdot)$ corresponds to a deterministic function mapping the hidden states $\boldsymbol{x}_{gt}$ to the same scale as the observed data y_{gt} (e.g. collapsing continuous event times to counts of events in the period $(t-1, t]$). (We note that $\boldsymbol{x}_{gt}$ might be of higher dimension than $\boldsymbol{y}_{gt}$, and thus $h_g(\cdot)$ might involve reducing the dimensionality of $\boldsymbol{x}_{gt}$, hence the slightly unwieldy notation.) The tolerances, ϵ_{gt}, define how closely we require the transformed hidden states $h_g(\boldsymbol{x}_{gt})$ to match the data y_{gt}. As $\epsilon_{gt} \to 0$ for all g and t, then the transformed hidden states are required to match the data exactly.

Particle filtering techniques aim to approximate (2) by using a finite set of *particles*, each corresponding to a particular realisation of the hidden states. In this context we could implement a classic bootstrap particle filter (Gordon et al., 1993), such that N particles are initialised by drawing independently from $f_{X_0}(\boldsymbol{x}_0 \mid \boldsymbol{\theta})$ and then propagated over time by simulating from the underlying model—with probability density $f_{X_t|X_{t-1}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta})$—and weighting each particle according to the discrepancy function $f_{Y_t|X_t,\epsilon_t}(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \epsilon_t)$. There are various straightforward ways to simulate from the kinds of stochastic compartmental models being studied in this paper, such as Gillespie's algorithm (Gillespie, 1977), and thus in the ABC setting described above, the problem can be reduced to simulating from the underlying model and recording whether the simulation is consistent with the data according to the discrepancy function.

A key feature of particle filters is that an unbiased, non-negative estimate of the desired likelihood, $\hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta})$, can then be obtained as a direct output from the filter. However, these filters often struggle with particle degradation, where many particles end up with zero (or close to zero) weights, which can cause instability in the likelihood

Algorithm 1 Alive Particle Filter.

Require: Number of particles N, parameters $\boldsymbol{\theta}$, and tolerances $\boldsymbol{\epsilon}_t$, for $t = 1, \ldots, T$.

Initialise the log of the estimated likelihood: $\log\left[\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}\right)\right] = T\log(N)$.
for $t = 1, \ldots, T$ **do**
  Set $n_t = 0$.
  for $n = 1, \ldots, N+1$ **do**
    Set $\delta = 0$.
    while $\delta = 0$ **do**
      if $t = 1$ **then**
        Set $r = n$ and sample initial states $\boldsymbol{x}_0^r$ from $f_{X_0}\left(\cdot \mid \boldsymbol{\theta}\right)$.
      else
        Sample an index r uniformly from the set $\{1, \ldots, N\}$.
      end if
      Simulate $\boldsymbol{x}_t'$ from $f_{X_t|X_{t-1}}\left(\cdot \mid \boldsymbol{x}_{t-1}^r, \boldsymbol{\theta}\right)$.
      Set $n_t = n_t + 1$.
      if $f_{Y_t|X_t}\left(\boldsymbol{y}_t \mid \boldsymbol{x}_t', \boldsymbol{\epsilon}_t\right) = 1$ **then**
        Set $\boldsymbol{x}_t^n = \boldsymbol{x}_t'$ and $\delta = 1$.
      end if
    end while
  end for
  Set $\log\left[\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}\right)\right] = \log\left[\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}\right)\right] - \log\left(n_t - 1\right)$.
end for

estimates. This is keenly felt when the discrepancy function produces binary weights, in which case it is possible to end up with all particles having exactly zero weight. Various extensions to the bootstrap particle filter have been proposed that aim to try to mediate these challenges, and the reader is encouraged to see e.g. Doucet et al. (2001) and Doucet and Johansen (2011) for comprehensive overviews of particle filtering techniques. Recently, Del Moral et al. (2015) proposed the *alive particle filter* (APF) as a means of tackling this problem for binary weights. The APF continuously resamples particles at each time point until N particles with non-zero weights have been obtained. This potentially solves the particle degradation problem, at the cost of making the number of simulations a random variable with a theoretically infinite range. In practice the filter can be stopped after a pre-defined number of simulations has been exceeded, improving the efficiency of particle MCMC algorithms at the cost of some bias in the estimated posteriors (see Drovandi et al., 2016). We return to this problem in Section 4.

The APF is outlined in Algorithm 1. The estimated likelihood (integrating over the hidden states) is

$$\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}\right) = \prod_{t=1}^{T} \frac{N}{n_t - 1}, \tag{5}$$

where N is the number of particles, and n_t is the number of simulations from the underlying model required to obtain $N+1$ matches in the period $(t-1, t]$ (this *must* be $N+1$ matches, in order to ensure an unbiased estimate—Del Moral et al., 2015).

Algorithm 2 ABC pseudo-marginal Metropolis-Hastings (ABC-PMMH) algorithm. As $\boldsymbol{\epsilon} \to \mathbf{0}$ then the algorithm converges to the true posterior in probability.

Require: Number of iterations M and tolerances $\boldsymbol{\epsilon}$.

1: Initialise parameters $\boldsymbol{\theta}^{(0)}$ and calculate an unbiased, non-negative estimate of the likelihood $\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}^{(0)}\right)$.
2: **for** $i = 1, \ldots, M$ **do**
3: Propose candidate parameters $\boldsymbol{\theta}'$ from some proposal distribution $q(\cdot)$ (here we use a random-walk proposal).
4: Calculate an unbiased, non-negative estimate of the likelihood $\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}'\right)$.
5: Calculate the acceptance probability:

$$\alpha = \min\left(1, \frac{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}'\right)}{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}^{(i-1)}\right)} \times \frac{f\left(\boldsymbol{\theta}'\right)}{f\left(\boldsymbol{\theta}^{(i-1)}\right)} \times \frac{q\left(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}'\right)}{q\left(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(i-1)}\right)}\right).$$

6: Sample $u \sim U(0, 1)$.
7: **if** $u < \alpha$ **then**
8: Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}'$.
9: **else**
10: Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$.
11: **end if**
12: **end for**

Andrieu and Roberts (2009) showed that substituting an *unbiased*, *non-negative* estimate of $f\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right)$ into a standard Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) will produce samples from the exact posterior in probability. This general *pseudo-marginal* Metropolis-Hastings algorithm is shown in Algorithm 2, and following Del Moral et al. (2015) and Drovandi et al. (2016) we use the APF to produce a suitable unbiased estimate $\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right)$. We will refer to Algorithm 2 as the ABC-PMMH algorithm.

2.2 Bayesian model choice

The ABC-PMMH algorithm provides a straightforward way to produce estimates of the (approximate) posterior distribution in the presence of hidden states. However, it is also often the case that there are various different ways in which the dynamics of the system could be modelled, which could lead to a very different understanding of how the underlying processes operate. Weighting the degree-of-evidence in favour of different competing models can help to elucidate some of these key mechanisms.

Bayesian model choice is frequently built around the concept of Bayes' Factors and posterior probabilities of association (Jeffreys, 1935, 1961). Formally, if we have W competing models to choose from—denoted $M_1, \ldots, M_W$—then we can derive a posterior probability for model M_w as

$$P\left(M_w \mid \boldsymbol{y}\right) = \frac{f\left(\boldsymbol{y} \mid M_w\right) P\left(M_w\right)}{\sum_{l=1}^{W} f\left(\boldsymbol{y} \mid M_l\right) P\left(M_l\right)}, \tag{6}$$

where $P(M_w)$ is the *prior* probability for model M_w. We note that this is not an absolute measure of model adequacy, rather it provides a *relative* posterior weighting conditional on the choice of W models. However, assuming no uncertainty in the choice of model can often lead to over-confident inferences (e.g. Kass and Raftery, 1995; Hoeting et al., 1999), and techniques such as such Bayesian Model Averaging (BMA), where applicable, can use these posterior weightings to help to improve the robustness of model predictions. The reader is referred to Kass and Raftery (1995) and Hoeting et al. (1999) for comprehensive introductions to these techniques.

Alternative Bayesian model choice approaches include the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), and posterior predictive p-values (e.g. Gelman et al., 2013). The former provides a Bayesian analogy to the popular Akaike's Information Criterion (AIC; Akaike, 1974), and is attractive because it can be easily calculated directly using posterior samples. It does not provide a probabilistic weight associated with each model, and thus cannot be used for model averaging, and is most useful for choosing between nested models. Posterior predictive p-values are a useful tool that can be used effectively to assess goodness-of-fit of competing models, which can then be used to inform model choice. The idea is to set up a loss function, which is evaluated over a set of posterior predictive samples to generate a measure of discrepancy against a pre-specified ideal scenario. An effective use of this approach is given in Lau et al. (2014), in which posterior predictive p-values are used to inform on the validity of different choices of distance kernel for a spatially explicit infectious disease model. However, as with DIC, these approaches do not provide a relative probabilistic weight, and thus cannot be used for model averaging.

As such, the focus of this manuscript is to estimate Bayes' Factors, and by extension posterior probabilities of association. The key quantity that underpins these ideas is the *marginal likelihood*:

$$f(\boldsymbol{y} \mid M_w) = \int_{\boldsymbol{\Theta}_w} f(\boldsymbol{y} \mid \boldsymbol{\theta}_w, M_w) f(\boldsymbol{\theta}_w \mid M_w) d\boldsymbol{\theta}_w, \tag{7}$$

where the integral is over the (multidimensional) parameter space for model M_w denoted by $\boldsymbol{\Theta}_w$. The marginal likelihood is equivalent to the normalising constant in (1) and is often referred to as the *Bayesian evidence*. In contrast to many frequentist quantities for weighting models, such as AIC, the Bayesian evidence *marginalises* (rather than maximises) over the parameter space. This means that uncertainties due to hidden states or auxiliary variables are explicitly incorporated into the Bayesian evidence, and that Bayes' Factor based model choice techniques are naturally parsimonious.

The key computational challenge is how to estimate the integral in (7), particularly in the presence of hidden states or missing information. We refer the reader to Touloupou et al. (2018) for a discussion of recent approaches to tackling this problem. In Touloupou et al. (2018), the authors provided an efficient and robust means of estimating the Bayesian evidence using importance sampling, and showed that it outperformed existing methods (see also Tran et al., 2013). Drovandi et al. (2016) implemented this approach explicitly in terms of using the APF developed for the ABC-PMMH algorithm as a means of integrating over the hidden states, and we extend these ideas here.

These techniques involve a two-stage process: firstly each competing model is fitted to the data using some numerical algorithm, such as MCMC, and then an approximation to the posterior distribution is used as an importance sampling distribution for estimating (7). To start with, consider that in the presence of hidden states the marginal likelihood for an arbitrary model (dropping the model index for brevity) can be written:

$$f(\boldsymbol{y} \mid \boldsymbol{\epsilon}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\epsilon}) f(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \, d\boldsymbol{x} \, d\boldsymbol{\theta}. \tag{8}$$

Under the paradigm of Wilkinson (2013), as $\boldsymbol{\epsilon} \to \mathbf{0}$ then the true marginal likelihood (for a model with no model discrepancy) is recovered. Drovandi et al. (2016) suggest approximating (8) as

$$\hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}) = \frac{1}{R} \sum_{r=1}^{R} \frac{\hat{f}(\boldsymbol{y} \mid \boldsymbol{\theta}_r) f(\boldsymbol{\theta}_r)}{q(\boldsymbol{\theta}_r)}, \tag{9}$$

where the $\boldsymbol{\theta}_r$ are sampled from a distribution with probability density function $q(\cdot)$, which is an approximation to the posterior distribution $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ and can be obtained by fitting an auxiliary distribution to the posterior samples obtained from the ABC-PMMH algorithm. Tran et al. (2013) show that (9) produces an unbiased estimate of $f(\boldsymbol{y} \mid \boldsymbol{\epsilon})$, and for a given $\boldsymbol{\theta}_r$, the estimate $\hat{f}(\boldsymbol{y} \mid \boldsymbol{\theta}_r)$ can be calculated using the APF (see also Touloupou et al., 2018; Alzahrani et al., 2018).

3 Data and model structures

The full data for the four experimental settings—as a set of time-series counts (and ignoring the swab information)—are shown in Figure 1. The initial experimental conditions are summarised in Table 1. We assume that all deaths are recorded at the end of each half-day period. We do not observe infection times, and so have removal information only. For simplicity we assume that culled moribund birds would have died naturally before the end of the corresponding time period in which they were euthanased (and are thus treated as natural removals within a given half-day period). We note that there are also healthy birds that were killed for immunohistological studies, which correspond to censored animals.

We will assume the system follows a stochastic, homogeneously mixing, frequency-dependent SIR model, in a closed population of size N. Hence at any point in time birds are either *susceptible* (S), *infected and infectious* (I), or *removed* (dead; R). We use superscripts to denote genotype, such that for a given experimental setting, there are N^{TG} *transgenic* birds and N^{NTG} *non-transgenic* birds (with total number of birds $N = N^{TG} + N^{NTG}$). We define a range of models, with the simplest model assuming that all challenge birds are infected on inoculation with probability p, and that there are no differences in either susceptibility to infection or onward transmission of infection between the genotypes. In this case we can characterise the probability of $S \to I$ or $I \to R$ moves in some small time interval $(t, t + dt)$ as:

$$\begin{aligned} P(S \to I \text{ in } (t, t+\delta t)) &= \frac{\beta SI}{N} dt + o(dt), \\ P(I \to R \text{ in } (t, t+\delta t)) &= \gamma I dt + o(dt), \end{aligned} \tag{10}$$

where $I = I^{TG} + I^{NTG}$ is the total number of infected birds, $S = S^{TG} + S^{NTG}$ is the total number of susceptible birds and $N = S + I$ is the total number of birds in the cage. Here β is the *transmission parameter*, and γ is the *removal rate* (hence $1/\gamma$ is the mean infectious period).

We can extend this model in various ways. Firstly, we could assume that the probability of infection post-inoculation varies between genotypes (e.g. with probability p^{TG} and p^{NTG} for the transgenic and non-transgenic birds respectively). We can also allow for different transmission terms (β^{TG} and β^{NTG}), different susceptibility terms (parameterised such that $\nu^{TG} = 1$ and $\nu^{NTG} \neq 1$), and/or different removal rates (γ^{TG} and γ^{NTG}). Hence the most complex model is characterised by:

$$\begin{aligned}
P\left(S^{NTG} \to I^{NTG} \text{ at time } 0\right) &\sim \text{Bin}\left(S^{NTG}, p^{NTG}\right), \\
P\left(S^{TG} \to I^{TG} \text{ at time } 0\right) &\sim \text{Bin}\left(S^{TG}, p^{TG}\right), \\
P\left(S^{NTG} \to I^{NTG} \text{ in } (t, t+\delta t)\right) &= \frac{\nu^{NTG} S^{NTG}}{N}\left(\beta^{TG} I^{TG} + \beta^{NTG} I^{NTG}\right) dt + o(dt), \\
P\left(I^{NTG} \to R^{NTG} \text{ in } (t, t+\delta t)\right) &= \gamma^{NTG} I^{NTG} dt + o(dt), \\
P\left(S^{TG} \to I^{TG} \text{ in } (t, t+\delta t)\right) &= \frac{S^{TG}}{N}\left(\beta^{TG} I^{TG} + \beta^{NTG} I^{NTG}\right) dt + o(dt), \\
P\left(I^{TG} \to R^{TG} \text{ in } (t, t+\delta t)\right) &= \gamma^{TG} I^{TG} dt + o(dt).
\end{aligned} \tag{11}$$

The full list of models considered are characterised in Table 2.

In this particular situation we have four independent experiments, and thus we can run the APF for each data set in turn, with an unbiased estimator of the likelihood being given by:

$$\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\epsilon}\right) = \prod_{k=1}^{K} \prod_{t=1}^{T_k} \frac{N_k}{n_{kt} - 1}, \tag{12}$$

where $k = 1, \ldots, K$ denotes the data sets (with $K = 4$ here), N_k is the number of particles used in the k^{th} particle filter, and n_{kt} is the number of simulations required to obtain $N_k + 1$ matches in time period $(t-1, t]$. Since there are two different genotypes (NTG and TG birds), the different experiments have different numbers of removal curves. Experiments 1 and 4 have a single curve each (for NTG and TG birds respectively), whereas experiments 2 and 3 both have two removal curves—one for each genotype. (In the parlance of Section 2.1, we have that $G = 1$ for experiments 1 and 4, and $G = 2$ for experiments 2 and 3.) We simplify matters by choosing a common tolerance, ϵ across all time-points and all curves. Therefore in experiments 1 and 4 a 'match' is obtained if the simulated number of removals matches the data shown in Figure 1, whereas for experiments 2 and 3 a 'match' is obtained if the simulated number of removals for *both genotypes* matches each corresponding curve in Figure 1 simultaneously—see equation (3).

Data augmentation and censoring

Although the APF will generate N particles with non-zero weight at each time point, we note an additional challenge for infectious disease models, which is that particles can

Model	Initial infection		Transmission		Susceptibility		Removal	
Model 1	p		β		1		γ	
Model 2	p^{NTG}	p^{TG}	β		1		γ	
Model 3	p		β^{NTG}	β^{TG}	1		γ	
Model 4	p^{NTG}	p^{TG}	β^{NTG}	β^{TG}	1		γ	
Model 5	p		β		ν^{NTG}	1	γ	
Model 6	p^{NTG}	p^{TG}	β		ν^{NTG}	1	γ	
Model 7	p		β^{NTG}	β^{TG}	ν^{NTG}	1	γ	
Model 8	p^{NTG}	p^{TG}	β^{NTG}	β^{TG}	ν^{NTG}	1	γ	
Model 9	p		β		1		γ^{NTG}	γ^{TG}
Model 10	p^{NTG}	p^{TG}	β		1		γ^{NTG}	γ^{TG}
Model 11	p		β^{NTG}	β^{TG}	1		γ^{NTG}	γ^{TG}
Model 12	p^{NTG}	p^{TG}	β^{NTG}	β^{TG}	1		γ^{NTG}	γ^{TG}
Model 13	p		β		ν^{NTG}	1	γ^{NTG}	γ^{TG}
Model 14	p^{NTG}	p^{TG}	β		ν^{NTG}	1	γ^{NTG}	γ^{TG}
Model 15	p		β^{NTG}	β^{TG}	ν^{NTG}	1	γ^{NTG}	γ^{TG}
Model 16	p^{NTG}	p^{TG}	β^{NTG}	β^{TG}	ν^{NTG}	1	γ^{NTG}	γ^{TG}

Table 2: Competing model specifications. Model 1 corresponds to the model defined in equation (10), where there are no differences between NTG and TG birds. Model 16 corresponds to the model defined in equation (11), where each component of the model (probability of infection following challenge, transmission, susceptibility and recovery) differ between NTG and TG birds. Intermediate models can be derived by equating different components between the genotypes (e.g. Model 15 can be derived from Model 16 by setting $p^{NTG} = p^{TG} = p$ and so on).

have non-zero weight at a time point t, but will never produce particles with non-zero weights at time points $> t$. This can happen, for example, if the number of infectives is zero at time t, but there are additional infections at a time point $t^* > t$. In the SIR framework, the probability of any further infection events is zero if the number of infectives $I = 0$. In this manuscript, we follow the approach of Drovandi et al. (2016) and augment the data at each time point to include the information regarding whether further infections happen at later time points. For $\epsilon = 0$, then this is equivalent to requiring that $I_t > 0$ at time point t if there are additional infections at later time points. For $\epsilon > 0$ we require that $I_t > 0$ if the cumulative number of infections $C_t^I < N_I - \epsilon$, where N_I is the maximum number of infections observed in the data. Simulations that do not adhere to these constraints are rejected. We only built in information on censoring for exact matching, at which point we randomly removed an animal at the end of the corresponding time period, and before the matching criteria was employed.

4 Improving the efficiency of the APF in ABC-PMMH algorithms

The main challenge when using the APF within an ABC-PMMH algorithm is that the number of simulations required to evaluate the APF is theoretically unbounded, and

thus could result in a huge computational burden, particularly since the filter needs to be run for a large number of MCMC iterations. This is further compounded when estimating the marginal likelihood where many more evaluations of the APF are required. Finally we must repeat this process for multiple models. However, in development terms the ABC-PMMH approach is attractive, since the method is fairly straightforward to code and adapt to different models. Therefore methods for improving the efficiency of the APF in the context of ABC-PMMH can help to alleviate the computational burden of the models and increase the size and complexities of systems that can be modelled using this technology.

One way to stop excessive runtimes in the APF is to put a cap on the total number of simulations that the filter is allowed to evaluate (see Drovandi et al., 2016). Therefore if, at some time point $j \leq T$,

$$\sum_{t=1}^{j} n_t > N^s, \tag{13}$$

then the estimate $\hat{f}(\boldsymbol{y} \mid \boldsymbol{\theta})$ is set to 0, where N^s is some arbitrary (large) integer. If implemented in an ABC-PMMH routine, then this would result in proposals being automatically rejected if the total number of simulations exceeds N^s. This introduces bias into the estimated posterior distribution, but the bias will decrease as $N^s \to \infty$. Thus choosing a suitable value for N^s is a trade-off between efficiency and bias; it needs to be large enough that few proposals result in N^s being exceeded, but small enough that it doesn't result in excessively long runtimes. In small systems this may not be an issue, but in larger systems, especially in poor regions of parameter space, or for a poorly specified model, then this can have much more of an impact on the efficiency of the ABC-PMMH routine. (Note that since the filtering is done over different time points, we define the total number of 'simulations' to aggregate over the time points, rather than a single simulation being a single realisation across all time points.)

We term the proportion of iterations where the proposal is rejected due to the particle filter failing to complete within N^s simulations (i.e. failing the criteria in 13) as the *skip rate*. A major challenge is that as the tolerances get small, the skip rate naturally increases, and this will induce more bias in the estimated posterior unless N^s is also increased. We now introduce an idea that aims to mediate this problem, by noting that it is possible to reject proposals without necessarily having to evaluate the APF to completion. This reduces both the computational load of the standard APF and the bias introduced by the cut-off N^s.

To formulate this idea, consider that at a given iteration i of the ABC-PMMH algorithm (Algorithm 2), we accept a proposal, $\boldsymbol{\theta}'$, with probability:

$$\alpha = \min\left(1, \frac{\hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}')\, f(\boldsymbol{\theta}')\, q\left(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}'\right)}{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}^{(i)}\right) f\left(\boldsymbol{\theta}^{(i)}\right) q\left(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(i)}\right)}\right). \tag{14}$$

Computationally, this is equivalent to simulating a random number $u \sim U(0,1)$, and then accepting $\boldsymbol{\theta}'$ if $u < \alpha$. Usually we calculate $\hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}')$, and thus α before we simulate u. However, since u is independent of α, there is no reason why we cannot simulate u first, meaning that the only unknown in equation (14) is $\hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}')$. Therefore, we

reject $\boldsymbol{\theta}'$ if:

$$u > \frac{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}'\right) f\left(\boldsymbol{\theta}'\right) q\left(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}'\right)}{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}^{(i)}\right) f\left(\boldsymbol{\theta}^{(i)}\right) q\left(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(i)}\right)}$$

$$\iff \quad \log\left[\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}'\right)\right] < \log(u) - A^{(i)}, \tag{15}$$

where $A^{(i)} = \log[f(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}')] - \log[\hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}, \boldsymbol{\theta}^{(i)})f(\boldsymbol{\theta}^{(i)})q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{(i)})]$. From (5) we can re-write the inequality (15) as

$$\sum_{t=1}^{T} \log\left(n_t - 1\right) > T\log(N) - \log(u) + A^{(i)}. \tag{16}$$

Now the only unknown is on the left-hand side of equation (16), and this is a monotonically increasing function with respect to n_t, where the sum must be conducted in order from $t = 1, \ldots, T$. Hence, for some time point $j \leq T$, if n_j reaches a point such that

$$\log\left(n_j - 1\right) > T\log(N) - \log(u) + A^{(i)} - \left[\sum_{t=1}^{j-1} \log\left(n_t - 1\right)\right], \tag{17}$$

then the proposal can be rejected with no loss-of-information. This places a finite upper bound on the number of simulations required to reject the proposal that depends only on the previous iteration and the simulation of the random variable u. If all time points evaluate such that $\sum_{t=1}^{T} \log(n_t - 1) < T\log(N) - \log(u) + A^{(i)}$ then the proposal is accepted.

We note that this idea to swap the order of evaluations in MCMC accept-reject steps has been employed before in other contexts, notably Beskos et al. (2006) and Papaspiliopoulos and Roberts (2008), who employed a similar approach in diffusion processes and Dirichlet processes respectively, to evaluate an infinite dimensional function in finite time. They termed their approaches "retrospective sampling". More recently, Solonen et al. (2012) used these ideas when evaluating the likelihood function in complex climate models, and termed their approach "early rejection sampling". This was extended into the ABC framework by Picchini (2014). Similar ideas of early stopping of simulations were used in early ABC papers such as McKinley et al. (2009), though these did not invert the accept-reject step, but exploited the evaluation of specific non-decreasing summary measures to stop simulations once it was known that it would be rejected. There are also links to ideas of squeezing approaches to rejection sampling (see e.g. Devroye, 1986), in the sense of rejecting proposals without having to evaluate complex functions to completion. However, these ideas have not been employed on the APF, and are important since the evaluation of the APF could take theoretically infinite time.

We also note some recent advances by Deligiannidis et al. (2018) and Tran et al. (2017), known as *correlated* and *block* pseudo-marginal methods respectively, along with the forward simulation algorithm of Neal and Huang (2015). These aim to improve computational efficiency by creating a non-centred parameterisation of the model—in the vein of Papaspiliopoulos et al. (2003)—and then storing the random numbers used to create the MC estimate of the likelihood. By correlating the random numbers used to generate the likelihood estimates for the current and proposed parameter values at

each iteration of the MCMC chain, then these approaches can be used to reduce the number of particles needed in order to obtain efficient mixing. Whilst these approaches look promising, we do not explore them further here due to the need to keep track of the random variables used for each simulation, regardless of whether or not the simulation is accepted, which can be prohibitively large. In addition, the correlated methods for updating random variables work best when the observation process is continuous, and in our case this is a binary indicator based on whether the simulations match the observed data according to a pre-defined tolerance. Thus small changes in the non-centred random variables can lead to a previously accepted simulation being rejected or visa-versa, which may affect the performance of the resulting MCMC algorithm.

4.1 Further efficiency savings

We note that for a given time period j, the value of $\log(n_j - 1)$ increases slowly with respect to increasing n_j. For example, for a fixed computational effort, $a+b$ simulations say (with $a > 1$ and $b > 1$), then $\log(a+b) \leq \log(a) + \log(b)$. Hence it makes sense to split computational effort across multiple time points if possible (since we are summing a set of logged counts), rather than focus on individual time points one at a time (where we are logging a single count). However, since the simulation order cannot be changed in the particle filter, we can instead try an alternative, but related approach, by considering writing (16) as:

$$\left[\sum_{t=1}^{j-1} \log(n_t - 1)\right] + \log(n_j - 1) + \left[\sum_{t=j+1}^{T} \log(n_t - 1)\right] > T\log(N) - \log(u) + A^{(i)}. \quad (18)$$

If we are currently evaluating time point j, then the term in the first set of square brackets in equation (18) has already been calculated. The term in the second set of square brackets in equation (18) has yet to be evaluated and so is ignored in terms of whether we reject the proposal based on the current value of n_j. However, we know that there is a minimum value that the future terms $\sum_{t=j+1}^{T} \log(n_t - 1)$ can take. This corresponds to every simulation at future time points being accepted first time. Under that scenario, $n_t = N+1$ for $t = j+1, \ldots, T$. Substituting this into (18) and re-arranging means that at time point j we can *reject* the proposal if

$$\log(n_j - 1) > T\log(N) - \log(u) + A^{(i)} - (T-j)\log(N) - \left[\sum_{t=1}^{j-1} \log(n_t - 1)\right]. \quad (19)$$

The term of the right-hand side of (19) is now smaller than the term of the right-hand side of (17), and thus we have reduced the upper bound on n_j.

This algorithm will converge with no bias in the resulting posterior distribution. Furthermore, it has the advantage that proposals are rejected more efficiently, requiring fewer simulations than would be required using the traditional approach. Although this approach provides an upper bound for the number of simulations required to reject a proposal, in practice, there are still situations where the absolute number of simulations

evaluated is still very high, and hence in practice it is still necessary to place some fixed upper bound N^s on the total number of simulations to reduce the overall computational burden. However, in general the proposed approach reduces both the skip rate *and* the computational load of the algorithms.

In Supplementary Section S1 (McKinley et al., 2019) we discuss additional extensions to this idea that worked well for our specific data. The first, which we denote APF*, extends the ideas outlined above to incorporate the fact that the data come from multiple independent experiments. The second, which we denote APF** exploits the independence of the different experiments to enable us to split the computational load across the different experiments in order to improve efficiency further.

These approaches only work as part of an MCMC algorithm, and for calculating the marginal likelihood in (9) one has to use the standard APF. We also note that the general idea of simulating u first could also be used in other routines, such as PMMH algorithms using the standard bootstrap filter, or even exact MCMC where the computation of the likelihood function is expensive. However, we expect the efficiency savings to be higher in the APF case, since there is greater variability in runtimes of the filter.

4.2 Efficiency of methods

We compare these approaches by fitting four different models (Models 1, 4, 7 and 16 from Table 2) to the Avian Influenza Virus (AIV) experimental data in Figure 1. We produced 15 runs of the ABC-PMMH algorithm, using the standard APF (with no efficiency adjustments), as well as the APF* and APF** approaches. Initial values were randomly sampled from the priors in each case, and each run consisted of 10,000 iterations burn-in, followed by a further 10,000 iterations. We used $N = 100$ particles and a tolerance of $\epsilon = 4$ for all data points, with the skip condition N^s set to $2{,}000 \times N = 200{,}000$ simulations. The results are shown in Figure 2, and the timings were taken from the final 10,000 iterations (to try to alleviate any bias that might be caused by individual chains being initialised far away from the area of high posterior mass).

We can see that both the APF* and APF** filters have a much lower skip rate in general than the standard APF. This corresponds to shorter run times. We note that the efficiency savings will depend greatly on the computational burden of the simulation model. Here the simulator is fast to evaluate, so the savings are not perhaps as pronounced as they would be for more computationally intensive models. In some cases the APF* approach was slightly quicker than the APF** approach, which results from the higher overheads of the APF** approach switching between the different time-series. However, we note that these differences are not large, and furthermore the APF** routine has a negligible skip rate in nearly all cases (for these settings), and so we think this represents the optimal option going forwards.

4.3 Using the APF for model choice using importance sampling

Once we have a set of posterior samples then we can use these to build a suitable proposal distribution to derive an importance sampling estimate of the marginal likelihood as

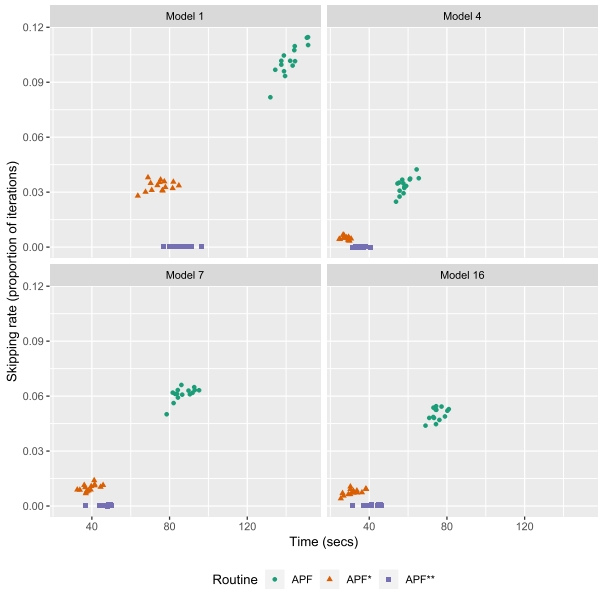


Figure 2: Comparative runs of the ABC-PMMH algorithm using each of the APF, APF* and APF** approaches.

given in (9). One challenge that arose during the simulation study was that poor models often had higher skip rates than good models. This results in biased posterior estimates from the ABC-PMMH algorithm. However, as long as the bias is not too great, then we can still obtain a suitable approximation to the posterior to use as a proposal distribution in the marginal likelihood calculation. Nonetheless, the requirement to use overdispersed proposal distributions in (9), meant that often skip rates during the marginal likelihood calculations (i.e. the proportion of the R proposals that were skipped) were higher than those obtained from the ABC-PMMH algorithm. This was exacerbated by the fact that since the R replicates are sampled independently it was necessary to use the standard APF for the marginal likelihood calculations. Hence an approach for dealing with skipped simulations is required.

One option is to set $\hat{f}(\boldsymbol{y} \mid \boldsymbol{\theta}_r) = 0$ for skipped proposals. However, this will underestimate the true marginal likelihood and lead to bias in the estimates. If the skip rate is not too high, then it may be that this bias is negligible, however for larger skip rates then the bias could become important. In principal these challenges could be tackled by increasing N^s for models with high skip rates. However, this can lead to high computational burdens, particularly for poor models or if the importance sampling distribution is too overdispersed. We decided to tackle this problem by finding upper and lower bounds for the estimate (9) in the presence of the skipped proposals, and then employing a conservative Occam's Window (e.g. Kass and Raftery, 1995) approach to remove poorly supported models. This latter approach has been justified in the literature as a means of removing models that have been scientifically discredited in the sense of having

little support under the data and priors compared to better models (e.g. Madigan and Raftery, 1994; Kass and Raftery, 1995).

Bounding the marginal likelihood estimates

We note that since the proposals are independent, then (9) can be written:

$$\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}\right) = \frac{1}{R}\left[\sum_{r=1}^{R_1} \frac{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\theta}_r\right) f\left(\boldsymbol{\theta}_r\right)}{q\left(\boldsymbol{\theta}_r\right)} + \sum_{r=R_1+1}^{R} \frac{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\theta}_r\right) f\left(\boldsymbol{\theta}_r\right)}{q\left(\boldsymbol{\theta}_r\right)}\right], \tag{20}$$

where indices $r = 1, \ldots, R_1$ correspond to the *non-skipped* proposals, and $r = R_1 + 1, \ldots, R$ to the *skipped* proposals. A lower bound for (9) can then be given by

$$\hat{f}_L\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}\right) = \frac{1}{R}\sum_{r=1}^{R_1} \frac{\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\theta}_r\right) f\left(\boldsymbol{\theta}_r\right)}{q\left(\boldsymbol{\theta}_r\right)}, \tag{21}$$

which is equivalent to setting the likelihood estimates for the skipped proposals to zero.

An upper bound for (9) can be found by considering that if the APFs for each of the K experiments are evaluated sequentially, then the log of the estimator (12) can be written:

$$\begin{aligned}\log\left[\hat{f}\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\epsilon}\right)\right] = &\left[\sum_{k=1}^{l-1}\sum_{t=1}^{j-1}\left\{\log\left(N_k\right) - \log\left(n_{kt} - 1\right)\right\}\right] \\ &+ \log\left(N_l\right) - \log\left(n_{lj} - 1\right) \\ &+ \left[\sum_{t=j+1}^{T_l}\left\{\log\left(N_l\right) - \log\left(n_{lt} - 1\right)\right\}\right] \\ &+ \left[\sum_{k=l+1}^{K}\sum_{t=1}^{T_k}\left\{\log\left(N_k\right) - \log\left(n_{kt} - 1\right)\right\}\right].\end{aligned} \tag{22}$$

Similarly to Supplementary Section S1 we note that if the proposal skips during time period j in experiment l, then the terms in the first line in (22) have already been evaluated; the second line has been partially evaluated; and the third and fourth line have yet to be evaluated. The maximum value that the third and fourth lines can take is zero, and the maximum value that the second line can take is $\log\left(N_l\right) - \log\left(n_{lj} + \left[N_l + 1 - n^*\right] - 1\right)$ where n^* is the number of particles with non-zero weight in the current time period j. This bound follows from the fact that we need to continue simulating until we obtain an extra $N_l + 1 - n^*$ matches, which would take at least $N_l + 1 - n^*$ additional simulations. Hence an upper bound for $\log[\hat{f}(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\epsilon})]$ can be given by:

$$\log\left[\hat{f}_U\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\epsilon}\right)\right] = \left[\sum_{k=1}^{l-1}\sum_{t=1}^{j-1}\left\{\log\left(N_k\right) - \log\left(n_{kt} - 1\right)\right\}\right] + \log\left(N_l\right) - \log\left(n_{lj} + N_l - n^*\right), \tag{23}$$

and therefore an upper bound for (9) can be given by:

$$\hat{f}_U(\boldsymbol{y} \mid \boldsymbol{\epsilon}) = \frac{1}{R}\left[\sum_{r=1}^{R_1} \frac{\hat{f}(\boldsymbol{y} \mid \boldsymbol{\theta}_r) f(\boldsymbol{\theta}_r)}{q(\boldsymbol{\theta}_r)} + \sum_{r=R_1+1}^{R} \frac{\hat{f}_U(\boldsymbol{y} \mid \boldsymbol{\theta}_r) f(\boldsymbol{\theta}_r)}{q(\boldsymbol{\theta}_r)}\right]. \tag{24}$$

Hence, $\hat{f}_L(\boldsymbol{y} \mid \boldsymbol{\epsilon}) \leq \hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon}) \leq \hat{f}_U(\boldsymbol{y} \mid \boldsymbol{\epsilon})$ where $\hat{f}_L(\boldsymbol{y} \mid \boldsymbol{\epsilon}) = \hat{f}_U(\boldsymbol{y} \mid \boldsymbol{\epsilon}) = \hat{f}(\boldsymbol{y} \mid \boldsymbol{\epsilon})$ when no proposals are skipped (i.e. when $R_1 = R$).

Problem specific extension

We can also perform a similar trick to Supplementary Section S1 where we split the computational effort amongst the different time series and iterate through them until we either complete the calculation or skip the proposal. In the ABC-PMMH algorithm the aim of this approach was to reduce the number of simulations required to reject proposals. In the marginal likelihood calculation the aim is instead to reduce the upper bound estimate $\hat{f}_U(\boldsymbol{y} \mid \boldsymbol{\epsilon})$.

Appealing to pragmatism

In practice the upper bounds $\hat{f}_U(\boldsymbol{y} \mid \boldsymbol{\epsilon})$ can be large, particularly when proposals are skipped at early points in the time series. This means that particles in poor regions of the parameter space can still require very large values of N^s in order to reduce the upper bound enough for us to make reasonable inference. For example, if we pick a particle with a very low transmission rate in a high transmission setting, then the probability of matching at each time point is very low, and thus we can use up our N^s simulations at the first couple of time points, but still have a high upper bound due to having evaluated only a few time points (since for a fixed computational effort, $a + b$ simulations say—with $a > 1$ and $b > 1$—then $\log(a + b) \leq \log(a) + \log(b)$). Given that the proposals are evaluated independently, if we saved the state of the system for the skipped proposals then we could restart the simulations with a higher N^s for the skipped particles. Here we chose a simpler solution, and just re-ran the skipped simulations with a higher N^s until the upper bound was sufficiently close to the lower bound.

There is also a relationship between the number of particles N used in the APF, and the variance of the importance sampling estimate (9), such that higher values of N required lower numbers of proposals R to get the same variance in the estimator. We chose to run the APF using the value of N optimised for the ABC-PMMH algorithm (see Supplementary Section S2) for each model, but note that it is not a requirement of the importance estimate to use a fixed N. For skipped proposals in the marginal likelihood calculation, we opted to reduce N to a single particle when re-running since this meant we could increase N^s to high values but control the computational burden for the subset of very poor proposals.

We generated estimates of uncertainty in the marginal likelihood estimates using bootstrapping. Hence we re-sampled (with replacement) the R samples used to estimate the marginal likelihood, and calculated the upper and lower bounds for each of these bootstrapped replicates. From these we could derive empirical confidence intervals for

the upper and lower bounds. Thus if the bootstrapped confidence intervals for the upper and lower bound estimates were similar then we would conclude that we didn't have to re-run any more skipped proposals; whereas the size of the confidence intervals help to inform about whether or not we need to increase the number of proposals R.

Sequential Occam's Window

To decide on how to remove models, we calculated the maximum lower bound estimate across the competing models:

$$\hat{f}_L^M\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}\right) = \max\left[\hat{f}_L\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, M_1\right), \ldots, \hat{f}_L\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, M_W\right)\right].$$

We then set the minimum difference between the best fit model for model w (on the log-scale) as:

$$\log\left[\hat{f}^*\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, M_w\right)\right] = \log\left[\hat{f}_L^M\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}\right)\right] - \log\left[\hat{f}_U\left(\boldsymbol{y} \mid \boldsymbol{\epsilon}, M_w\right)\right],$$

before selecting all models w such that $\log[\hat{f}^*(\boldsymbol{y} \mid \boldsymbol{\epsilon}, M_w)] \leq \log(20)$. This will provide a conservative way to select models that reduces to an exact Occam's Window criterion when the skip rate for all models is zero.

The key point here is that we may not have to increase N^s for poor models, as long as the upper bound is sufficiently far away from the best model, so this method allows us a pragmatic way to understand the impact of the skip rate on the marginal likelihood estimates. We found that once Occam's Window has been employed, then the remaining models tended to fit the data better and thus had lower skip rates, whereas poor models took a disproportionate share of the computational burden and often ended up with negligible posterior weight. Since both the number of simulations and the skip rate can be reduced by using larger tolerances, we chose to employ a sequential Occam's Window approach (across a series of waves) to efficiently remove poor models from the analysis. Initially, we performed model selection at some value ϵ^{target}, chosen such that the models were required to fit the data reasonably well, but so that the skip rate for the worst models was not too high. We then removed models where the marginal likelihood was less than a factor of 20 away from the best model using the conservative Occam's Window routine described previously. We then repeated the whole fitting process again, using only those models selected at the previous wave, with starting tolerances ϵ^{ini} equal to the original target tolerance and a new target tolerance $\epsilon^{\text{target}} \leq \epsilon^{\text{ini}}$. The cut-off N^s can be tweaked at each stage to reflect that the matching probability decreases as the tolerance decreases. This is not a perfect solution to the problem, since the *ratio* of marginal likelihoods for two models at a tolerance of ϵ is not always larger than the same ratio at a tolerance of $\epsilon^* < \epsilon$, hence there is a chance that we remove a model at an earlier stage that may have been included at a later stage. However, as long as the initial target tolerance is small enough that there is enough information in the approximate posterior to distinguish between relatively good models and the set of poor models, then it is unlikely to make a huge impact, since the models that were removed would most likely have had low posterior weights in any case. (We choose a factor of 20 as

our cut-off following Kass and Raftery, 1995, however this could be made more or less stringent if desired.)

We note that this approach improves efficiency greatly, since the computational load tends to be spread disproportionately across poor models, given that the probability that they produce a match is much lower and thus they require a much larger number of simulations and/or particles in order to evaluate. Removing these models earlier allows us to focus computational effort on models that are more likely to produce good fits to the data and thus reduces the computational burden.

The complete training pipeline that we used is described in detail in the Supplementary Section S2. We optimised the number of particles N at each stage using the approach of Sherlock et al. (2015), and used an adaptive MCMC proposal distribution from Roberts and Rosenthal (2009)—see Supplementary Section S3.

5 Simulation study

To test the performance of the algorithms we conducted a simulation study. We picked a series of models in Table 2 (Models 1, 4, 7, 16) and a set of parameter values (listed in Supplementary Table S1). For each model we ran 1,000 replicate simulations. Hence for each time point we obtained a discrete distribution of counts, from which we could estimate the mode count. We then picked the individual simulation that was closest to the mode (according to its L^2 norm across all time points). These provided a series of simulated data sets on which we could test our algorithms. All simulations were aggregated to counts of infection and removals at daily time intervals.

In Section 5.1 we explore these approaches in large population sizes (scaling the cohorts in Lyall et al., 2011 by 4, providing 20 challenge birds and 48 in-contact birds in each experiment). We then ran the simulated experiments for 30 days. This was in order to test the scalability of these techniques in larger populations and their performance in data-rich situations. Section 5.2 then provides analogous simulation studies where the experimental setup mirrors that of Lyall et al. (2011) (shown in Table 1).

For each model we chose uniform $U(0.01, 1)$ priors for any parameters bounded between zero and one, and gamma Exp(1) priors for any other parameters. In all cases we used the same tolerance, ϵ, around each data point. We ran each simulation study twice, the first time requiring that simulations matched both infection *and* removal curves within each experiment, and the second time requiring that simulations matched just the removal curves.

5.1 Large-population simulations

Pseudo-code for the training pipeline described in Section 4.3 is given in Algorithm S1. To summarise: in the first wave we start by producing a short training run of M^{train} MCMC iterations using a "large" initial tolerance ϵ^{ini} and an initial number of particles N. From this training run we extract the posterior medians for each of the parameters, and use these to optimise the number of particles N. (Pseudo-code for this optimisation

procedure is provided in Algorithm S2.) We repeat this process multiple times using smaller tolerances each time until we hit some value ϵ^{target}. At this point we then produce longer MCMC runs for each model, from which we can calculate the marginal likelihood estimates and apply the conservative Occam's Window approach described in the previous section to remove poor models. We repeat this whole procedure for a series of subsequent waves, using the remaining models from the previous wave, with the initial tolerance ϵ^{ini} set equal to ϵ^{target} from the previous wave, before updating $\epsilon^{\text{target}} < \epsilon^{\text{ini}}$.

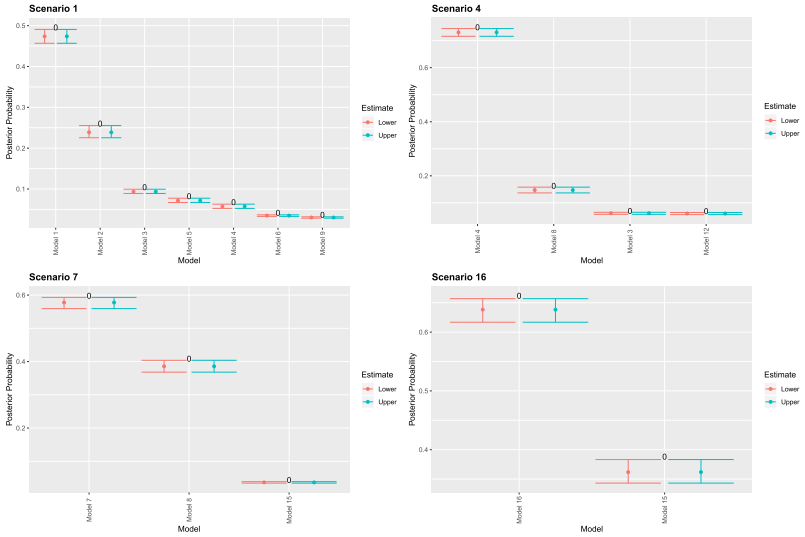
For each simulation scenario we ran a single initial wave at a fixed threshold of $\epsilon^{\text{ini}} = 20$, starting with $N = 100$ particles. We sampled initial values, $\boldsymbol{\theta}^{\text{ini}}$, from the prior(s). Since the optimal number of particles, N, varied between the different models, we chose to define the skip cut-off, N^s, as a multiple of the number of particles, such that $N^s = 10{,}000 \times N$. In practice we found that these default choices were usually sufficient to provide good mixing, although occasionally we tweaked them if we couldn't find suitable starting values.

At each stage we used two chains of $M^{\text{train}} = 10{,}000$ iterations for each training run, and reduced the tolerance such that $\epsilon \mapsto \epsilon - 1$ each time. As described in Supplementary Section S2 we sometimes needed to extend the training runs for longer depending on the mixing. We used $N^{\text{rep}} = 500$ replicates when optimising the number of particles (see Algorithm S2 for details), and chose default ranges of between 1 and 200 particles across which to optimise. After training, the final run produced two chains of 50,000 (approximate) posterior samples assuming that $\epsilon^{\text{target}} = 20$.

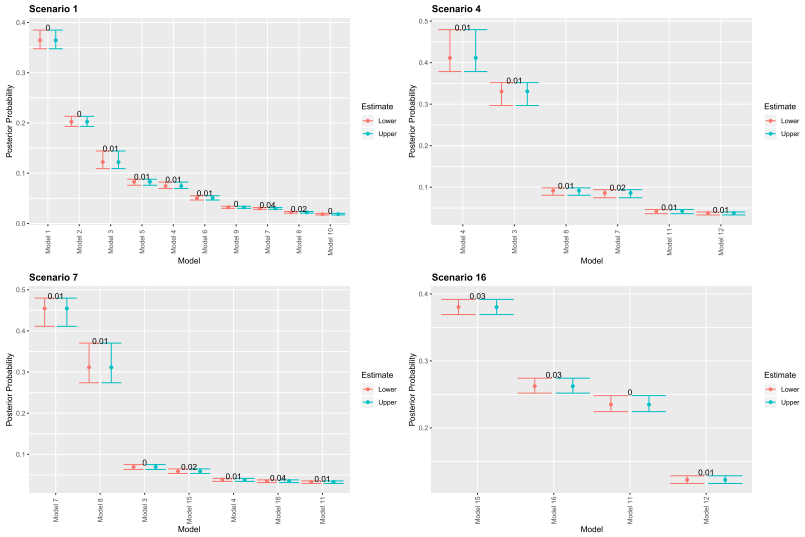
We then used an approximation to this posterior distribution from this longer run to calculate the marginal likelihood estimates for each model using the method described in Section 4.3, with $R = 20{,}000$ samples. The proposal distribution was chosen to be a truncated multivariate normal distribution, with a covariance matrix equal to the empirical covariance matrix of the posterior samples, but with the diagonals scaled by 1.05 (to make the proposals overdispersed compared to the posterior). The truncation was necessary to ensure that the proposals were on the correct scale (bounded in $(0, 1)$ for proportions, or > 0 for all other parameters). We chose prior probabilities for each model to be equal, such that $P(M_w) = 1/W$, where W is the number of competing models. To get a feel for the uncertainty in the estimates, we generated 95% confidence intervals for the posterior probabilities and marginal likelihoods using 100 bootstrapped replicates.

Subsequent waves assumed values of $\epsilon^{\text{ini}} = 19$ and $\epsilon^{\text{target}} = 12$, followed by $\epsilon^{\text{ini}} = 11$ and $\epsilon^{\text{target}} = 5$, and finally $\epsilon^{\text{ini}} = 4$ and $\epsilon^{\text{target}} = 1$. The simulated removal curves, along with the target regions defined by $\epsilon^{\text{target}} = 20,\ 12,\ 5$ and 1 are shown in Supplementary Figure S1. For brevity we do not show the infection curves. When it was necessary to re-run the skipped simulations during the marginal likelihood calculations, we used $N = 1$ particle and a cut-off $N^s > 1{,}000{,}000$ for the re-runs.

We can see the log-marginal likelihood estimates from the intermediate waves in Supplementary Figure S2 for the case where we match to both infection and removal curves; and Supplementary Figure S4 for the case where we match to the removal curves only. For brevity we show the posterior weights for each remaining model from the final waves in Figure 3. We can see that in each case the routine picks the correct model,



(a) Fit to infection and removal curves



(b) Fit to removal curves only

Figure 3: Posterior probability estimates at the final wave for the large-scale simulation studies, for all scenarios across all simulation variants: (a) fitting to infection and removal curves; and (b) fitting to removal curves only. Uncertainty bounds come from 100 bootstrapped replicates.

with larger weights on the correct model in the case where we match to both infection and removal information, relative to the case where we match to removal curves only. The only exception is for Scenario 16 when fitting to removal curves only. In this case there is preferential support for the simpler M_{15} over M_{16}, which corresponds to the loss-of-information due to matching to removal curves only.

The corresponding weighted posterior distributions for the parameters are shown in Supplementary Figures S3 and S5, and again we see an improvement in inference if we include more information in the data, though in all cases the approximate posteriors capture the true values for the parameters. One point to note is that when we have set the probability of initial infection to 1, such as for the p^{NTG} parameter in scenarios 4 and 16, then this value sits on the upper bound of the prior distribution, and hence by definition the bulk of posterior mass lies below this point. We chose to stop at $\epsilon = 1$, and hence the degree-of-approximation to the exact posteriors should be small here.

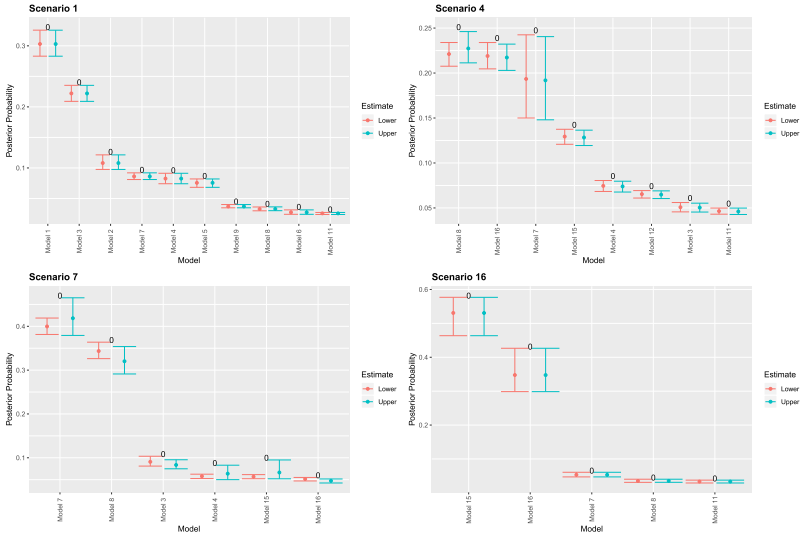
5.2 Small-population simulations

In the small-scale case we used a similar approach to Section 5.1, using an initial tolerance of $\epsilon^{\text{ini}} = 5$ and a target tolerance of $\epsilon^{\text{target}} = 2$ in the first wave, with $\epsilon^{\text{ini}} = 1$ and $\epsilon^{\text{target}} = 0$ in the final wave. The simulated removal curves, along with the target regions defined by $\epsilon^{\text{target}} = 2$ and 0 are shown in Supplementary Figure S6. Again, for brevity we do not show the infection curves.

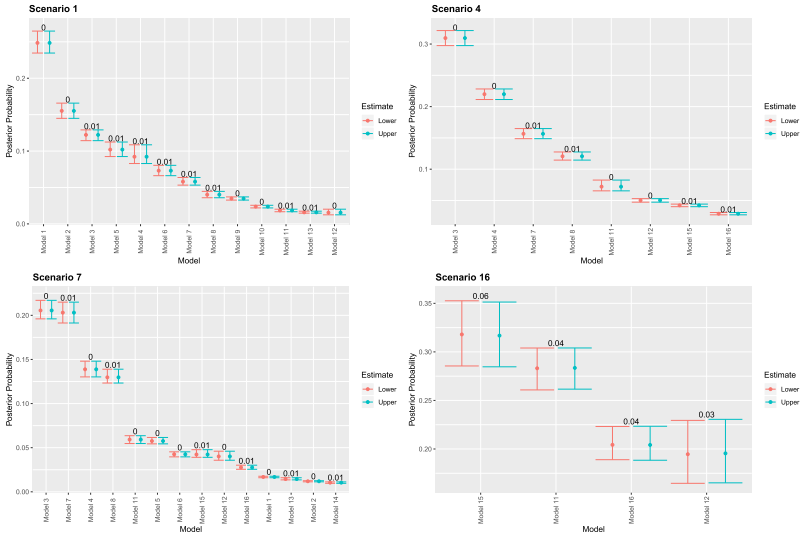
We can see the log-marginal likelihood estimates from the intermediate waves in Supplementary Figure S7 for the case where we match to both infection and removal curves, and Supplementary Figure S9 for the case where we match to the removal curves only. For brevity we show the posterior weights for each remaining model from the final waves in Figure 4.

In contrast to the larger study, although the true model is always contained in the subset of final models, it is not always associated with the largest posterior weight. This is because there is much less information in the data, due to smaller sample sizes. In this case the Bayesian model choice paradigm will tend to naturally favour more parsimonious models where the model fits are comparable, though in some cases a lack of detailed information in the data can allow parameters to trade-off against each other. (This can be seen for Scenario 4 in Figure 4a, where the preferred model is M_8 rather than M_4. In this case the difference between these two models is simply that M_8 has an additional susceptibility term, which in this case allows it to fit the particular data set better than the simpler model.)

The corresponding weighted posterior distributions for the parameters are shown in Supplementary Figures S8 and S10, and again we see an improvement in inference if we include more information in the data, though in all cases the approximate posteriors adequately capture the true values for the parameters. We note that since $\epsilon = 0$ here, the final weighted posteriors and marginal likelihoods are estimates of the *exact* quantities.



(a) Fit to infection and removal curves



(b) Fit to removal curves only

Figure 4: Posterior probability estimates at the final wave for the small-scale simulation studies, for all scenarios across all simulation variants: (a) fitting to infection and removal curves; and (b) fitting to removal curves only. Uncertainty bounds come from 100 bootstrapped replicates.

Parameter	Mean
β^{NTG}	1.45
β^{TG}	0.48
γ^{NTG}	0.53
γ^{TG}	0.51
p^{NTG}	0.49
p^{TG}	0.50
ν^{NTG}	1.05

Table 3: Model averaged posterior summaries from the models fitted to the experimental data.

Parameter	Number of parameters	Posterior probability
β	1	0.21
	2	**0.79**
γ	**1**	**0.72**
	2	0.28
ν	**0**	**0.7**
	1	0.3
p	**1**	**0.58**
	2	0.42

Table 4: Model averaged posterior weights for the number of parameters of each type, from the models fitted to the experimental transmission data. Bold font highlights the specification with highest posterior support.

6 Application to experimental transmission study of AIV

We now apply our approach to the experimental transmission data from Lyall et al. (2011), shown in Figure 1 and Table 1. These data are openly available from the University of Exeter's institutional repository at: https://doi.org/10.24378/exe.1644. (Note that the measurements are now at half-day intervals, rather then daily intervals as per the simulation study.)

We used the same approach as in the Section 5, using an initial tolerance of $\epsilon^{\text{ini}} = 5$ and a target tolerance of $\epsilon^{\text{target}} = 2$ in the first wave, and a fixed tolerance of $\epsilon^{\text{ini}} = \epsilon^{\text{target}} = 1$ in the second wave. The target regions defined by $\epsilon^{\text{target}} = 2$ and 1 are shown in Supplementary Figure S11.

We can see the log-marginal likelihood estimates from the intermediate waves in Supplementary Figure S12. Weighted posterior distributions at the final wave are shown in Supplementary Figure S13 and weighted posterior means for each parameter are given in Table 3. Note that we can also derive model averaged posterior weights regarding certain hypotheses of interest, such as whether there are one or two transmission, susceptibility or initial infection probability terms. These are shown in Table 4. We can see that under these choices of models and priors, there is some evidence for a difference in transmission parameters between the competing models, but weaker evidence for differences between the other terms.

7 Discussion

We have presented a re-analysis of the experimental transmission study of Lyall et al. (2011) in which we provide estimates of the Bayesian evidence between a series of competing mechanistic models of transmission, up to an ABC tolerance of $\epsilon = 1$ around each data point. In this case, after the models with extremely poor support were removed (according to the Occam's Window criteria), there remained some evidence for a difference in the transmission parameters between the two genotypes (posterior probability of 0.79), but weaker evidence to support any difference in susceptibilities (posterior probability of 0.3). There was weak support for a difference in the probabilities of initial infection (posterior probability of 0.42). This reinforces the view in Lyall et al. (2011) that the observed differences in the removal curves were most likely being driven by differences in *transmission potential* between the genotypes, rather than differences in *susceptibility*. Furthermore, our methodology allows us to produce model averaged posterior distributions for the different epidemiological parameters, and in particular the analysis suggests that the non-transgenic birds show a much higher transmission potential than the transgenic birds (on average 3.0 times larger). We also note that we can use information about where the models struggle to fit to help diagnose model/data mismatches. For example, when we attempted to fit using an ABC tolerance of zero, the models became very computationally expensive. By monitoring where the routines skipped out, it became clear that most of the routines were struggling to capture the sharp rise in removals between days 6–6.5 in Group 2 (Figure 1). In fact, we made an assumption here that culled moribund birds would have died before the end of time period in which they were culled, which may have led to a slightly sharper spike at that time point than in the raw data (one bird was culled during this period in the TG class). This suggests that it may be beneficial to come up with a better way to handle these culled animals in future studies.

We performed a detailed set of simulation studies to assess the efficacy of the methodology. For large sample sizes the approach picked up the correct model in each case. For the scenarios where time series counts for both the infection and removal curves was known, then the weight of evidence for the 'true' model was stronger than the respective scenarios where the data were restricted to removal times only. For smaller sample sizes, where there is less information in the data to distinguish between the models, then the selection routines tend to pick up a larger set of supported models (as expected under the Bayesian paradigm).

We have made some contributions to the way in which ABC-PMMH algorithms can be implemented in order to reduce bias and improve efficiency when using computationally demanding estimators such as the APF. One challenge that remains is to develop more sophisticated ways to deal with skipped proposals when calculating the marginal likelihood estimates. In the examples in this paper, when we re-ran these simulations using a larger cut-off N^s, the upper bounds for the log-marginal likelihoods converged on the lower bounds. This means that parameters that produced estimates that skipped at early time points did end up having low posterior weights. However, it sometimes took very many more simulations to get to this point (usually setting $N = 1$ particle with N^s between 1,000,000 and 10,000,000 for the re-runs worked sufficiently well, though some proposals required an $N^s = 100{,}000{,}000$). For completeness we ran

these to completion, but in practice an improved rule-of-thumb for setting a suitable N^s would help to alleviate some of these challenges.

Another area that could be improved is the training of the filters. For consistency we used sequential ABC-PMMH runs, with different choices of tolerance ϵ as a means of calibrating our training runs. An alternative could be to use some more generic ABC sampler, such as ABC Sequential Monte Carlo (ABC-SMC; e.g. Toni et al., 2009) in the early stages, to cut down the prior space and produce suitable training data to help optimise the number of particles required for the alive or bootstrap particle filter-based MCMC. However, these still require the development of suitable summary statistics and tolerances. A related approach to the one presented here was developed in Drovandi and McCutchan (2016), in which the APF was used to generate estimates of the likelihood function that could be used in an (ABC-)SMC routine. They call their approach "alive SMC^2". This approach generates estimates of the marginal likelihood as a by-product of the fitting mechanism, rather than requiring a separate importance sampling step. However, the authors note that in some cases the SMC estimator of the marginal likelihood could contain a large amount of Monte Carlo variability, in which case a separate importance sample estimate such as the type employed in this manuscript can alleviate this. This is in line with Touloupou et al. (2018), who show that the importance sampling approach often outperforms various existing methods in terms of reducing the MC error of the marginal likelihood estimates. Similarly to mixing problems in pseudo-marginal MCMC caused by the use of an estimate in place of the true likelihood, Drovandi and McCutchan (2016) also note that if the MC error arising from the APF is too high, then they observe an increase in particle degradation, resulting in more re-sampling and perturbation steps being required. In addition, the alive SMC^2 algorithm will still suffer with the problems of high skip rates for poor models, and thus some of the ideas introduced here, relating to the sequential Occam's Window selection approaches may still be relevant.

We note alternative recent advances in model comparison (Pudlo et al., 2016) and model inference (Raynal et al., 2017) techniques that use random forests (Breiman, 2001) within the ABC paradigm. These approaches look promising, since they inform on optimal choices of summary statistics and do not require the specification of tolerances. They do require suitably large reference tables to be generated, and it is of great interest in future work to compare the performance of these approaches to the types of routine that have been developed in this manuscript. Another alternative might be to use a tempering approach (e.g. Ratmann et al., 2007).

Alternative estimation methods, such as constrained simulation algorithms (McKinley et al., 2014) could also be used within an MCMC algorithm to provide exact inference. However, these are much more challenging to implement and generalise to multiple models. This is true also of data-augmented (reversible-jump) MCMC (e.g. Gibson and Renshaw, 1998; O'Neill and Roberts, 1999), since again these methods are much harder to implement, even for a single model, let alone multiple models. In addition some form of forward simulation algorithm would still be required in order to generate the importance sample estimates of the marginal likelihoods. The advantage of the bootstrap or alive particle filters is that they are straightforward to code, and thus can be extended in

a straightforward manner to fit and compare different models, which is a key advantage when modelling complex dynamical systems.

Supplementary Material

Supplementary Materials: Efficient Bayesian model choice for partially observed processes: with application to an experimental transmission study of an infectious disease (DOI: 10.1214/19-BA1174SUPP; .pdf).

References

Akaike, H. (1974). "A new look at statistical model identification." *IEEE Transactions on Automatic Control*, AU-19: 195–223. MR0423716. doi: https://doi.org/10.1109/tac.1974.1100705. 8

Alzahrani, N., Neal, P., Spencer, S. E. F., McKinley, T. J., and Touloupou, P. (2018). "Model selection for time series of count data." *Computational Statistics and Data Analysis*, 122: 33–44. MR3765813. doi: https://doi.org/10.1016/j.csda.2018.01.002. 4, 9

Anderson, R. and May, R. (1991). *Infectious Diseases of Humans*. Oxford University Press. 3

Andrieu, C., Doucet, A., and Holenstein, R. (2010). "Particle Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society, Series B (Methodological)*, 72(3): 269–342. MR2758115. doi: https://doi.org/10.1111/j.1467-9868.2009.00736.x. 3, 4

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo simulation." *The Annals of Statistics*, 37(2): 697–725. MR2502648. doi: https://doi.org/10.1214/07-AOS574. 3, 4, 6

Beaumont, M. A. (2003). "Estimation of population growth and decline in genetically monitored populations." *Genetics*, 164: 1139–1160. 3, 4

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2006). "Retrospective exact simulation of diffusion sample paths with applications." *Bernoulli*, 12(6): 1077–1098. MR2274855. doi: https://doi.org/10.3150/bj/1165269151. 13

Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1): 5–32. MR3874153. 4, 27

Del Moral, P., Jasra, A., Lee, A., Yau, C., and Zhang, X. (2015). "The Alive Particle Filter and Its Use in Particle Markov Chain Monte Carlo." *Stochastic Analysis and Applications*, 33(6): 943–974. MR3415229. doi: https://doi.org/10.1080/07362994.2015.1060892. 4, 6, 7

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018). "The correlated pseudomarginal method." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 839–870. MR3874301. doi: https://doi.org/10.1111/rssb.12280. 13

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag New York. MR0836973. doi: https://doi.org/10.1007/978-1-4613-8643-8. 13

Doucet, A., de Freitas, N., and Gordon, N. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*. Springer. MR1847783. doi: https://doi.org/10.1007/978-1-4757-3437-9. 4, 6

Doucet, A. and Johansen, A. M. (2011). *A Tutorial on Particle Filtering and Smoothing: Fifteen years later*. Oxford University Press. MR2884612. 6

Drovandi, C. C. and McCutchan, R. A. (2016). "Alive SMC^2: Bayesian Model Selection for Low-Count Time Series Models with Intractable Likelihoods." *Biometrics*, 72: 344–353. MR3515761. doi: https://doi.org/10.1111/biom.12449. 27

Drovandi, C. C., Pettitt, A. N., and McCutchan, R. A. (2016). "Exact and approximate Bayesian inference for low integer-valued time series models with intractable likelihoods." *Bayesian Analysis*, 11(2): 325–352. MR3471993. doi: https://doi.org/10.1214/15-BA950. 4, 6, 7, 8, 9, 11, 12

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition. MR3235677. 8

Gibson, G. J. and Renshaw, E. (1998). "Estimating parameters in stochastic compartmental models using Markov chain methods." *IMA Journal of Mathematics Applied in Medicine and Biology*, 15: 19–40. 4, 27

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996). *Markov Chain Monte Carlo In Practice*. Chapman and Hall. MR1397966. doi: https://doi.org/10.1007/978-1-4899-4485-6. 4

Gillespie, D. T. (1977). "Exact stochastic simulation of coupled chemical reactions." *The Journal of Physical Chemistry*, 81(25): 2340–2361. 5

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." *Radar and Signal Processing, IEE Proceedings F.*, 140(2): 107–113. MR1229888. 5

Hastings, W. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57: 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 7

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: A tutorial." *Statistical Science*, 14(4): 382–417. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 8

Jeffreys, H. (1935). "Some tests of significance, treated by the theory of probability." *Proceedings of the Cambridge Philosophy Society*, 31: 203–222. 7

Jeffreys, H. (1961). *The Theory of Probability*. Oxford, 3rd edition. 7

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). "Bayesian analy-

sis for emerging infectious diseases." *Bayesian Analysis*, 4(4): 465–496. MR2551042. doi: https://doi.org/10.1214/09-BA417. 4

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 3, 8, 16, 17, 20

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. MR2354763. 3

Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. J. (2014). "New model diagnostics for spatio-temporal systems in epidemiology and ecology." *Journal of the Royal Society Interface*, 11: e20131093. 8

Lyall, J., Irvine, R. M., Sherman, A., McKinley, T. J., Núñez, A., Purdie, A., Outtrim, L., Brown, I. H., Rolleston-Smith, G., Sang, H., and Tiley, L. (2011). "Suppression of avian influenza transmission in genetically modified chickens." *Science*, 331(6014): 223–226. 2, 3, 20, 25, 26, 32

Madigan, D. and Raftery, A. E. (1994). "Model selection and accounting for model uncertainty in graphical models using Occam's Window." *Journal of the American Statistical Association*, 89: 1535–1546. 17

Marin, J.-M., Pillai, N. S., Robert, C. P., and Rousseau, J. (2014). "Relevant statistics for Bayesian model choice." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5): 833–859. MR3271169. doi: https://doi.org/10.1111/rssb.12056. 4

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). "Markov chain Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences USA*, 100(26): 15324–15328. 3, 4

McKinley, T. J., Cook, A. R., and Deardon, R. (2009). "Inference in epidemic models without likelihoods." *The International Journal of Biostatistics*, 5(1). MR2533810. doi: https://doi.org/10.2202/1557-4679.1171. 4, 13

McKinley, T. J., Neal, P., Spencer, S. E. F., Conlan, A. J. K., and Tiley, L. (2019). "Supplementary Materials: Efficient Bayesian model choice for partially observed processes: with application to an experimental transmission study of an infectious disease." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1174SUPP. 15

McKinley, T. J., Ross, J. V., Deardon, R., and Cook, A. R. (2014). "Simulation-based Bayesian inference for epidemic models." *Computational Statistics and Data Analysis*, 71: 434–447. MR3131981. doi: https://doi.org/10.1016/j.csda.2012.12.012. 27

McKinley, T. J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G. (2018). "Approximate Bayesian Computation and simulation-based inference for complex stochastic epidemic models." *Statistical Science*, 33(1): 4–18. MR3757500. doi: https://doi.org/10.1214/17-STS618. 4

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). "Equa-

tions of state calculations by fast computing machine." *Journal of Chemical Physics*, 21: 1087–1091. 7

Neal, P. and Huang, C. L. T. (2015). "Forward simulation Markov Chain Monte Carlo with applications to stochastic epidemic models." *Scandinavian Journal of Statistics*, 42: 378–396. MR3345110. doi: https://doi.org/10.1111/sjos.12111. 13

O'Neill, P. D. and Roberts, G. O. (1999). "Bayesian inference for partially observed stochastic epidemics." *Journal of the Royal Statistical Society. Series A (General)*, 162: 121–129. 4, 27

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 13

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). *Non-centered parameterizations for hierarchical models and data augmentation*, 307–326. Oxford University Press. MR2003180. 13

Picchini, U. (2014). "Inference for SDE Models via Approximate Bayesian Computation." *Journal of Computational and Graphical Statistics*, 23(4): 1080–1100. MR3270712. doi: https://doi.org/10.1080/10618600.2013.866048. 13

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M., and Robert, C. P. (2016). "Reliable ABC model choice via random forests." *Bioinformatics*, 32(6): 859–866. MR3304913. doi: https://doi.org/10.1007/s11222-014-9530-9. 4, 27

Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., and Wiuf, C. (2007). "Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*." *PLoS Computational Biology*, 3(11): e230. MR2369270. doi: https://doi.org/10.1371/journal.pcbi.0030230. 27

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., and Estoup, A. (2017). "ABC random forests for Bayesian parameter inference." URL https://arxiv.org/pdf/1605.05537. MR3889283. 4, 27

Roberts, G. O. and Rosenthal, J. S. (2009). "Examples of adaptive MCMC." *Journal of Computational and Graphical Statistics*, 18(2): 349–367. MR2749836. doi: https://doi.org/10.1198/jcgs.2009.06134. 20

Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). "On the efficiency of pseudo-marginal random walk Metropolis algorithms." *The Annals of Statistics*, 43(1): 238–275. MR3285606. doi: https://doi.org/10.1214/14-AOS1278. 20

Sisson, S., Fan, Y., and Tanaka, M. M. (2007). "Sequential Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences USA*, 104: 1760–1765. MR2301870. doi: https://doi.org/10.1073/pnas.0607208104. 3

Solonen, A., Ollinaho, P., Laine, M., Haario, H., Tamminen, J., and Järvinen, H. (2012). "Efficient MCMC for Climate Model Parameter Estimation: Parallel Adaptive Chains and Early Rejection." *Bayesian Analysis*, 7(3): 715–736. MR2981633. doi: https://doi.org/10.1214/12-BA724. 13

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society. Series B (Methodological)*, 64(4): 583–639. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 8

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997). "Inferring coalescence times from DNA sequence data." *Genetics*, 145: 505–518. 3, 4

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Strumpf, M. P. (2009). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems." *Journal of the Royal Society Interface*, 6: 187–202. 3, 4, 27

Touloupou, P., Alzahrani, N., Neal, P., Spencer, S. E. F., and McKinley, T. J. (2018). "Model comparison with missing data using MCMC and importance sampling." *Bayesian Analysis*, 13(2): 437–459. MR3780430. doi: https://doi.org/10.1214/17-BA1057. 8, 9, 27

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2017). "The block pseudo-marginal sampler." Technical report. URL https://arXiv:1603.02485v5. 13

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2013). "Importance sampling squared for Bayesian inference in latent variable models." URL https://arxiv.org/abs/1309.3339v3. 8, 9

Wilkinson, R. D. (2013). "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error." *Statistical Approaches in Genetics and Molecular Biology*, 12(2): 129–141. MR3071024. doi: https://doi.org/10.1515/sagmb-2013-0010. 4, 9

Acknowledgments

SEFS gratefully acknowledges funding by MRC grant MR/P026400/1 and EPSRC grant EP/R018561/1. The original experimental work of Lyall et al. (2011) was supported by the Biotechnology and Biological Sciences Research Council (grants BB/G00479X/1, BBS/B/00239, and BBS/B/00301) and by the Cambridge Infectious Diseases Consortium [Department for Environment, Food and Rural Affairs / Higher Education Funding Council for England—grant VT0105]. AJKC was supported by BBSRC grant BB/I024550/1.